CU-HTK STT Systems for RT03

Phil Woodland, Gunnar Evermann, Mark Gales, Thomas Hain,
Ricky Chan, Bin Jia, Do Yeong Kim, Andrew Liu,
David Mrva, Dan Povey, Khe Chai Sim, Marcus Tomalin
Sue Tranter, Lan Wang & Kai Yu

May 19th 2003

Cambridge University Engineering Department

Rich Transcription Workshop 2003

Presentation Overview

- Introduction
- Work on English CTS (Woodland)
- Development work on English Broadcast News (Kim)
- Fast System Descriptions (Evermann)
- Mandarin CTS (Woodland)
- Conclusions

2003 CU-HTK English CTS Systems

Phil Woodland, Ricky Chan, Gunnar Evermann, Mark Gales,
Thomas Hain, Do Yeong Kim, Andrew Liu, David Mrva,
Dan Povey, Sue Tranter, Lan Wang, Kai Yu

May 19th 2003

Cambridge University Engineering Department

English CTS Development

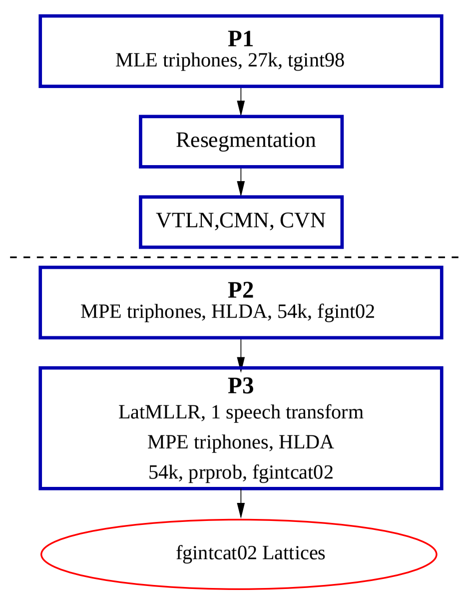
- 2002 unlimited computation system
- Training and test data sets
- New/Revised components
  - automatic segmentation
  - revised transcriptions
  - variable number of Gaussians
  - lattice generation for MPE training
  - SAT experiments
  - additional acoustic training data
  - SPron experiments
  - revised language models
- 2003 system performance
- Conclusions

2002 System

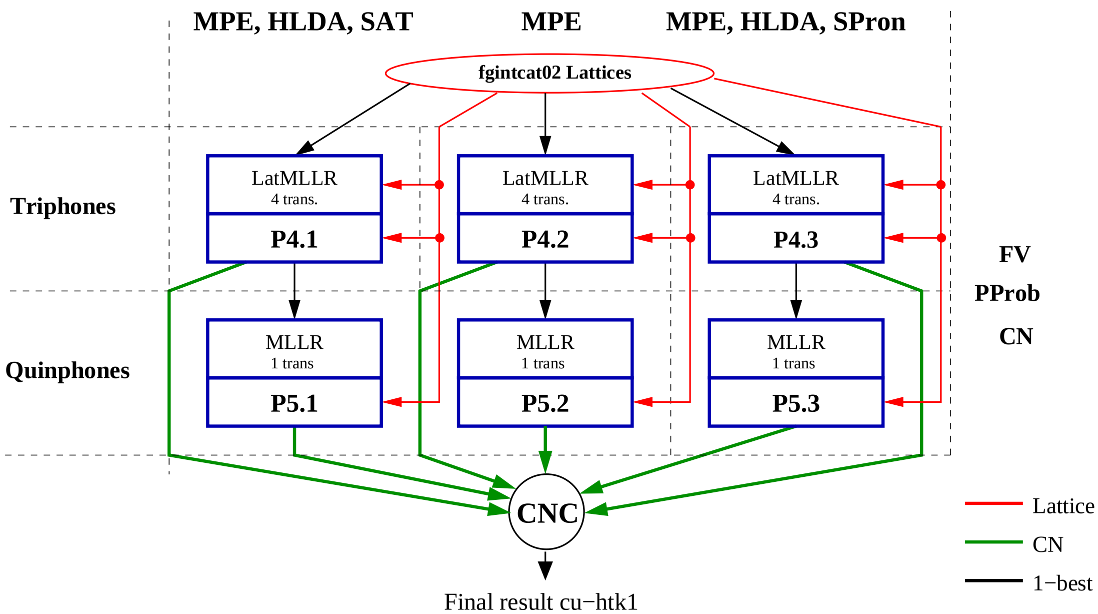
- Assumes manual segmentation into turns
- PLP, side-based CMN/CVN + 1st/2nd Δs (+ 3rd Δs & HLDA to 39 dims)
- Initial passes generate transcriptions for VTLN & initial adaptation
- Generates lattices with adapted triphone models and a bigram LM
- Expands the lattices to 4-gram plus trigram category model
- Rescores the lattices with adapted triphone and quinphone models
  - MPron HLDA SAT MPE triphone/quinphones
  - SPron HLDA non-SAT MPE triphones/quinphones
  - MPron non-HLDA non-SAT MPE triphones/quinphones
- Use confusion networks to represent each rescoring pass output & confusion network combination for highest posterior prob words and confidence scores

2002 System - Lattice Generation

- MLE P1 models
- MPE triphone models for P2/P3
- 28 mixture components (28 mix)
- HLDA
- Adaptation for P3 via Lattice MLLR
- Pronunciation probabilities
- HTK decoder HDecode



2002 system – Rescoring & Combination



Results on eval02 set

		Swbd1	Swbd2	Cellular	Total
P1	trans for VTLN	35.6	44.6	50.5	44.0
P2	trans for MLLR	24.6	30.9	34.8	30.4
P3	lat gen	22.5	28.0	31.3	27.5
P4.1	SAT tri	21.6	26.3	29.6	26.1
P4.2	non-HLDA tri	22.3	27.4	31.2	27.2
P4.3	SPron tri	21.5	26.6	29.1	26.0
P5.1	SAT quin	21.5	25.5	28.6	25.4
P5.2	non-HLDA quin	22.4	26.7	30.7	26.9
P5.3	SPron quin	21.5	26.4	28.8	25.8
CNC	P4.[123]+P5.[123]	19.8	24.3	27.0	23.9

%WER on eval02 for all stages of 2002 system, manual segmentation

- final confidence scores have NCE 0.289

Training and Test Data Sets

h5train02 248 hrs Switchboard (Swbd1), 17 hrs CallHome English (CHE) + LDC cell1 corpus (without dev01/eval01 sides) extra 17 hrs of data

h5train03 290 hr set. As above plus extra 12 hours of Switchboard I from final MSU transcripts

h5train03b 360 hr set. As above plus extra Switchboard Celluar I and Swd2 Phase2 data as released by BBN (CTRANS transcribed)

Development test sets

dev01 40 sides Swbd2 (eval98), 40 sides Swbd1 (eval00), 38 sides Swbd2 cellular (for manual segments)

eval02 40 sides of Swbd2; 40 sides of Swbd1; 40 sides of Swbd cellular. Can be used with manual or automatic segments

Automatic Segmentation

- Need to automatically segment the input data this year
- Used models with Gaussian mixture modes specific for cellular/non-cellular & male/female (256 Gaussians for male/female; 128 for silence)
- Constrained to have only one type of speech per side
- More details in diarisation talk

	Diarisation score (dryrun data)	% WER (eval02)
CUED dryrun segments	13.09	27.8
CUED sys03 segments	8.55	27.3
STM segments	39.89 (!)	26.7

Recogniser used in 10xRT system from Dec'02 (dryrun)

Revised Transcriptions

A mistake in the Switchboard training transcriptions used in building all CUHTK CTS systems since 2000 was discovered.

- Error in processing MSU Swbd training transcripts
- Some fairly common words systematically deleted (3% of tokens)
- Affected both acoustic models and LMs
- Rebuilt transcriptions based on final version of MSU transcripts
- Added 294 new conversation sides
- Rebuilding acoustic models only, for 2002 10xRT system on eval02 (manual segs), reduced WER by (only?) 0.5% abs (27.2 to 26.7)

Var #Gauss per state

- CU's std approach was N Gaussians per speech state and $2N$ for silence
- Set #Gauss as a function of number of frames γ_j available to train state j
- Use #Gauss $= k\gamma_j^p$, where p is a small power (e.g. 1/5)
- k is a normalising constant set to make the average #Gauss equal to N
- On CTS typically gives a 0.1-0.4% abs reduction in WER (see later tables)

SAT/Adaptation Experiments

- SAT tries to remove inter-speaker variability in training set by means of linear transform
- Use constrained MLLR to generate a single transform per training side (can operate in feature space)
- Interleave update of adaptation matrices and MLE HMM updates
- Perform MPE training based on SAT models with fixed transforms
- 0.3% abs improvement from SAT

	SAT				non-SAT			
	Sw1	Sw2	Cell	Tot	Sw1	Sw2	Cell	Tot
1 best std MLLR	17.7	31.1	30.5	26.4	17.7	31.6	31.0	26.7
lattice MLLR/FV	17.4	30.4	29.7	25.8	17.5	30.9	30.2	26.1

% WER dev01 manual seg 2002 fgintcat LM, HLDA MPE-trained triphones

Lattice-Based MPE Training

- Minimum Phone Error training (Povey & Woodland, 2002)
- Uses lattice-based training developed for MMI and extended B-W updates
- Includes "I-smoothing" of discriminative statistics with ML counts
- Requires the generation of lattices for the training set:
  - Correct transcription (corresponds to MMI numerator)
  - Representation of the confusable model sequences (MMI denominator)
- Denominator lattices generated in two steps
  - Word level lattice generation (uses training-data bigram LM)
  - Model-marking of HMM sequence and segmentation points (unigram LM)
  - Training procedure treats segmentation points as truth
  - Lattices generated using ML models (non-HLDA)

Modified Lattice-Based Training

- In 2002 no re-alignment/regeneration of lattices during discriminative training
  - In 2001 re-generated model-marked lattices part way through MMI training
- Now use heavily pruned training data bigram for word lattice generation
  - larger "denominator" lattices
  - better representation of confusable data
  - use pruned bigram scores in MPE training also
- Use HLDA ML models to generate lattices (rather than non-HLDA lattices)
- After 4 iterations of MPE training regenerate word and model-marked lattices with MPE models and use *both* of lattices (combining at statistics level).

MPE Training with Modified Lattices: Results

	Swbd1	Swbd2	Cellular	Total
non-HLDA lattices	20.5	35.3	34.7	30.1
HLDA full bg + ug	20.4	34.7	34.3	29.7
HLDA pruned bg	20.0	34.4	34.0	29.4
MPElattice regen/comb	19.4	34.0	33.6	28.9

% WER dev01 manual seg 2002 trigram LM, unadapted 28mix HLDA triphones, 290hr training, MPron

- HLDA ML models to generate lattices reduces WER by about 0.4% abs
- Larger lattices with pruned bigram reduce WERs by about 0.3% abs
- This lattice regen/comb gives a further 0.5% abs improvement in WER

Additional Acoustic Training

- New Swbd2 data transcriptions provided by BBN (70 hours)
- About 1% abs reduction in WER for MLE HMMs and 1.3% for MPE
- Largest improvement for cellular data (2.2% abs) and Swbd2 data (1.4% abs)

	290hr train				360hr train			
	Sw1	Sw2	Cell	Tot	Sw1	Sw2	Cell	Tot
16 comp MLE	24.9	39.8	39.6	34.7	24.7	39.4	38.5	34.1
28 comp MLE	24.0	39.0	38.1	33.6	23.6	38.1	36.8	32.7
Var comp (28) MLE	23.9	38.8	38.0	33.5	23.1	37.8	36.8	32.5
MPE (8its)	20.0	34.4	34.0	29.4	19.4	33.2	31.7	28.0
MPE lat combine	19.4	34.0	33.6	28.9	19.0	32.6	31.4	27.6

% WER dev01 manual seg 2002 trigram LM, unadapted HLDA triphones

SPron Dictionary

- Modified procedure from 2002 CUHTK CTS eval system (Hain, 2002)
- Systematically remove all pronunciation variants
- If words were observed in the training data
  - Selection is based on pronunciation variant frequency
  - DP alignment of pronunciation variant pairs followed by merging variants with substitutions only and then phoneme deletions/insertions
- Training of statistical model on decisions above
  - For a pair of pronunciation variants identify target and source
  - Model uses phoneme substitution probs
- Unobserved words
  - Identify source variant from statistical model
  - Select primary variant by pairwise exclusion

SPron experiments

- Rebuilt SPron models with MPE lattice comb from MPron word lattices
- Lattice combination helps 0.8% with SPron models built like this
- Final MPron and SPron WERs very similar (SPron 1% abs better for MLE)

	MPron				SPron			
	Sw1	Sw2	Cell	Tot	Sw1	Sw2	Cell	Tot
16 comp MLE	24.7	39.4	38.5	34.1	24.4	38.3	37.5	33.3
28 comp MLE	23.6	38.1	36.8	32.7	23.0	36.9	36.1	31.9
Var comp (28) MLE	23.1	37.8	36.8	32.5	22.6	36.6	35.6	31.5
MPE (8its)	19.4	33.2	31.7	28.0	19.9	33.1	32.2	28.3
MPE lat combine	19.0	32.6	31.4	27.6	19.0	32.2	31.6	27.5

% WER dev01 manual seg 2002 trigram LM, 360hr training, unadapted HLDA triphones

- After eval03 found SPron lattices from scratch (new word lattices/model marked lattice + MPE regen) helps by only another 0.1% absolute

2003 language models

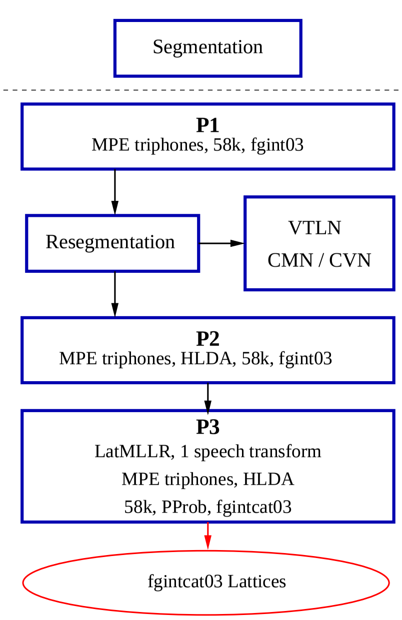
- Training data in 5 portions:
  - Revised MSU transcripts + CHE [3MW]
  - broadcast news setup (BN transcripts from PSM; CNN data; TDT data) [427MW]
  - Cell1 transcriptions [0.2MW]
  - Swb2 transcriptions from BBN/CTRANS [0.9MW]
  - google data from U of Washington [62MW]
- Used dev01, eval01 and eval02 as dev set
- Selected 30k words from acoustic transcripts plus top 54k words from BN (58k total). OOV rate 0.19% on dev set
- Trained 5 component 4-gram LMs; one class 3-gram LM

- “Small” text sources trained using modified Kneser-Ney (SRI LM); large text source using Good-Turing (HTK HLM)
- 2003 merged fgintcat has 4.3% rel reduction in PP over 2002 model. With cat models the difference is 3.5%
- Effect of component 4-gram word LMs

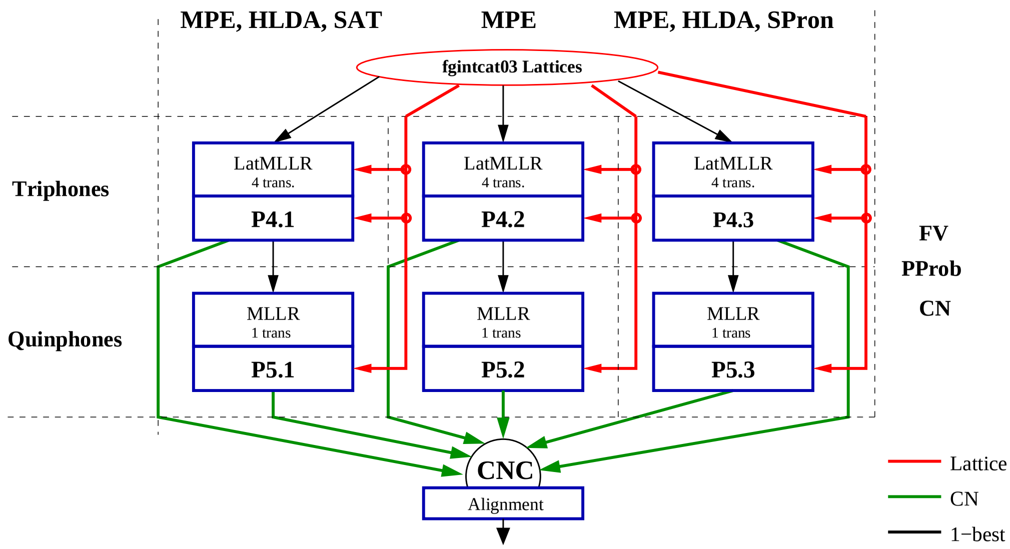
component LMs	fg PP
all	65.2
all minus google	65.9
all minus cell1	67.4
all minus swbdII	68.4
all minus che+swbdI	68.6
all minus BN+TDT+CNN	68.9

2003 System

- Automatic Segmentation
- Revised non-VTLN HTLDA MPE P1 models (290hr) + fg LM
- Revised MPE training for all other models (360hr)
- Modified SPron models for tri/quin
- Pronunciation probabilities in tri/quin
- Adaptation & system combination same
- Final alignment step



2003 System Part II



2003 System Performance (Eval02)

		Swbd1	Swbd2P3	Cellular	Total
P1	trans for VTLN	27.2	34.8	39.5	34.2
P2	trans for MLLR	23.6	28.9	31.7	28.4
P3	lat gen	21.1	25.1	27.6	24.8
P4.1	SAT tri	19.9	23.3	25.2	23.0
P4.2	non-HLDA tri	21.2	24.9	27.7	24.8
P4.3	SPron tri	20.4	23.7	25.6	23.4
P5.1	SAT quin	20.0	23.6	25.0	23.0
P5.2	non-HLDA quin	21.2	24.9	27.1	24.6
P5.3	SPron quin	20.1	23.9	25.3	23.3
CNC	P4.[123]+P5.[123]	18.6	22.3	23.7	21.7

%WER on eval02 for all stages of 2003 system (auto-segments)

Final NCE is 0.304

2003 System Performance (Eval03)

		Swbd2P5	Fisher	Total
P1	trans for VTLN	37.7	27.9	33.0
P2	trans for MLLR	31.8	22.6	27.4
P3	lat gen	27.5	19.3	23.5
P4.1	SAT tri	25.4	18.2	21.9
P4.2	non-HLDA tri	27.4	19.6	23.7
P4.3	SPron tri	25.6	18.5	22.2
P5.1	SAT quin	25.5	18.4	22.1
P5.2	non-HLDA quin	27.5	19.6	23.7
P5.3	SPron quin	25.7	18.7	22.3
CNC	P4.[123]+P5.[123]	24.1	17.1	20.7

%WER on eval03 (current test) for all stages of 2003 system (auto-segments)

Final NCE is 0.318

Conclusions

A number of changes and improvements have been made to the system although basic structure the same as 2002 system

- Automatic segmentation now gives only 0.6% increase in WER
- On eval02 data got 23.9% WER in 2002 with manual segments: now 21.7% with automatic segments. Approx 12% reduction in WER if use consistent manual segments
- revised Swb1 transcriptions: 0.5% abs
- variable number of Gaussians per state: 0.3% abs
- new MPE lattice generation/regeneration procedure: 1.2% abs
- new Swb2 data: 1.3% abs unadapted / no system combination

- revised language models: (guess) 0.2% abs on eval02 but expect more on Fisher data?

- Overall the system ran in 187xRT

- For Current Test Set error rate is 20.7%

- For Progress Test Set (all Fisher) error rate is 17.4%

- Many interesting things didn't make eval system this year
  - MMI/MPE training of HLDA transforms
  - discriminative estimation of SAT transforms
  - more advanced covariance modelling (e.g. extended MLLT)
  - ...

2003 CU-HTK Broadcast News English System Development

Do Yeong Kim, Gunnar Evermann, Thomas Hain,
David Mrva, Sue Tranter, Lan Wang, Phil Woodland,
and Rest of the HTK STT team

May 19th 2003

Cambridge University Engineering Department

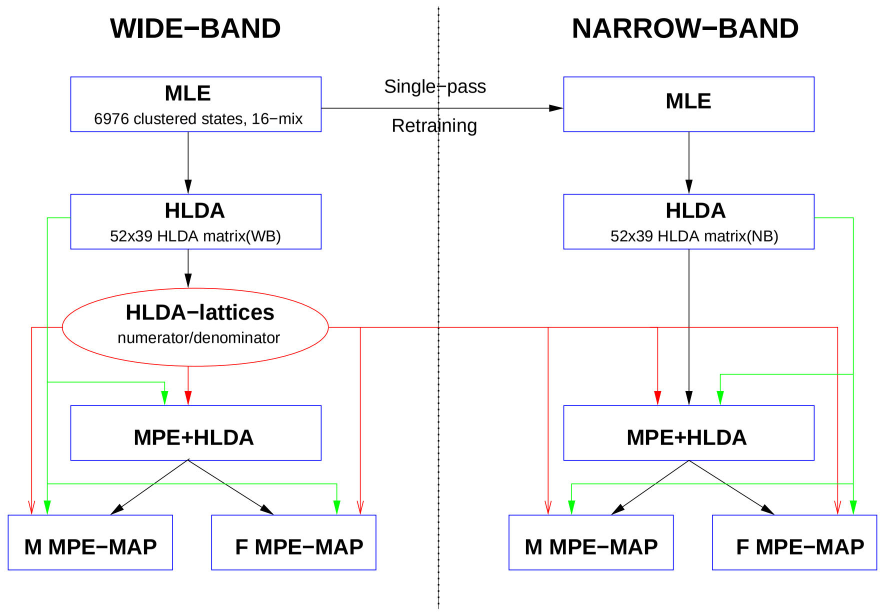
Overview

- Training data + Baseline Acoustic Models
- Adaptation Experiments
- Language Models
- Improved Acoustic Models
  - VarMix
  - Lattice-Regeneration MPE
- SAT
- SPron

Training data + Baseline Acoustic Models

- Training data : the 143 hours combined set of 1997 and 1998 data from LDC
  - **1997 data** 72 hours of acoustic BN training data
  - **1998 data** 71 hours of acoustic BN training data
- Front-end
  - 12 MF-PLP cepstral parameters + C0 and 1st/2nd derivatives + segment CMN (no VTLN or CVN)
  - Optional 3rd derivatives + HLDA
- Acoustic modelling
  - Decision tree state clustered, context dependent triphone models (6976 clustered states, 16-component mixture Gaussian)
  - Gender-dependent & bandwidth-dependent acoustic modelling
  - MLE/MPE/MPE-MAP training

Baseline Acoustic Models: Building Overview



Baseline Acoustic Models: Results (I)

- Development test sets

  BNeval98 two 1.5-hour data sets
  BNeval02 1-hour data set
  BNdev03 three hours of TDT-4 data from Jan '01 transcribed by STT sites

- 1998 CU-HTK BN-E LM (trigram)

- Single pass decoding without any adaptation

Baseline Acoustic Models: Results (II)

- HLDA transform
  - Estimate HLDA transform based on MLE baseline system
  - Add 3rd derivatives + HLDA, project 52 dim to 39
  - Consistent gain over different test sets, genders, and F-conditions
- MPE+HLDA
  - MPE training based on HLDA models
  - Significant gain over MPE or HLDA

	MLE	HLDA	MPE +HLDA
BNeval98			
F0	11.1	10.2	8.8
F1	20.1	18.5	15.5
F2	25.8	22.6	19.6
F3	20.9	19.1	17.3
F4	20.1	18.9	15.3
F5	28.1	27.2	19.1
FX	35.0	30.5	25.7
All	19.6	17.9	15.0
BNeval02			
All	17.9	16.0	13.6

%WER on BNeval98 & BNeval02

Basic Acoustic Models: MPE-MAP

- Gender-dependent discriminative training with MPE-MAP
  - Simple gender-dependent MPE model showed small gain (14.8%WER on BNeval98)
  - MAP-style update without losing advantage of discriminative training, see [Povey, Gales, Woodland: ICASSP2003]
- Most gains come from female speakers while both genders were improved

	MPE	MPE -MAP
BNeval98		
F	15.1	14.0
M	14.3	14.3
All	15.0	14.5
BNeval02		
F	14.8	14.5
M	13.3	12.5
All	13.6	13.0

%WER of MPE-MAP

Adaptation Experiments

	BNeval98			BNeval02		
	M	F	Total	M	F	Total
GI(HLDA+MPE)	14.3	15.1	15.0	12.9	15.3	13.6
1-best MLLR	13.8	14.4	14.4	12.0	14.1	12.6
Lat-MLLR 2trans	13.4	14.2	14.0	11.9	14.3	12.5
Lat-MLLR 2trans+FV	13.3	13.9	13.9	11.8	14.0	12.4
Lat-MLLR 4trans+FV	13.3	13.7	13.8	11.7	13.8	12.3

%WER for BNeval98 & BNeval02 after adaptation based on the GI unadapted models

- Apply global 1-best MLLR, phone-mark lattices, perform 4 iterations of Lattice MLLR
- By adapatation, WER was reduced by 8.7% relative on BNeval98, and 9.6% on BNeval02
- Small gains from FV and beyond 2 transforms

Improved Acoustic Model: Variable # of Gaussians

	BNeval98			BNeval02		
	F	M	Total	F	M	Total
HLDA	18.2	17.1	17.9	18.4	15.1	16.0
HLDA+VarMix	18.0	16.8	17.6	18.2	15.0	15.8

%WER on BNeval98 & BNeval02.

- Different number of Gaussians were assigned to each states according to the amount of available training data, while maintaining the average number of Gaussians per states the same as basic set-up (16 Gaussian/state)
- Marginal but consistent gains over two different test sets and both genders

Improved Acoustic Model: Lattice-Regeneration MPE

- Lattices for MPE training were regenerated using 4 iterations MPE+HLDA models with pruned bigram
- 4 more iterations of MPE with pruned bigram lattices and original lattices

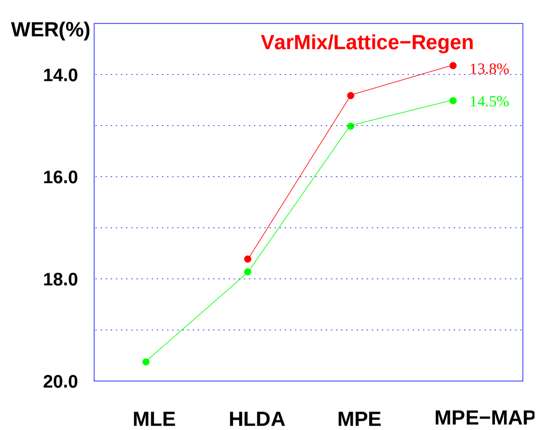
	Total	F0	F1	F2	F3	F4	F5	FX	F	M
MPE+HLDA	15.0	8.8	15.5	19.6	17.3	15.3	19.1	25.7	15.1	14.3
Lattice-Regen	14.4	8.5	15.1	17.7	16.9	14.6	21.3	24.4	14.5	13.8

%WER of Lattice-Regeneration MPE on BNeval98

- Lattice-Regeneration MPE reduced 0.6% abs. error rates, and outperformed MPE+HLDA models in almost every F-conditions except F5 (speech from non-native speakers).
- Also works with gender dependent models (0.7% abs gain)

Improved Acoustic Model: Summary

- VarMix showed marginal gain
- VarMix/Lattice-Regeneration significantly recuded WER both in MPE(GI) and MPE-MAP(GD)
- 29.6% of relative reduction in WER (5.8% abs.) on BNeval98 by progress in acoustic modeling



%WER on BNeval98

Language Model (I)

- Language model training texts: 1,019 MW in total
- Subsets for interpolation

	Source	epoch	size (MW)
A	Primary Source Media BN transcriptions	1992-1999	275
	TDT 2 & 3 closed captions		
B	CNN shows transcription	1999-2001	66
C	TDT4 closed captions		2
D	broadcast news acoustic training transcriptions	1997-1998	2
	acoustic transcriptions for Marketplace shows	1996	
E	Los Angeles Times newswire service texts	1995-1998	674
	Washington Post newswire service texts	1995-1998	
	New York Times newswire texts	1997-2001	

No data from dates after mid January 2001 was used to conform with the epoch restriction for the eval data (Feb. 2001) and the BNdev03 set (late Jan. 2001)

Language Model (II)

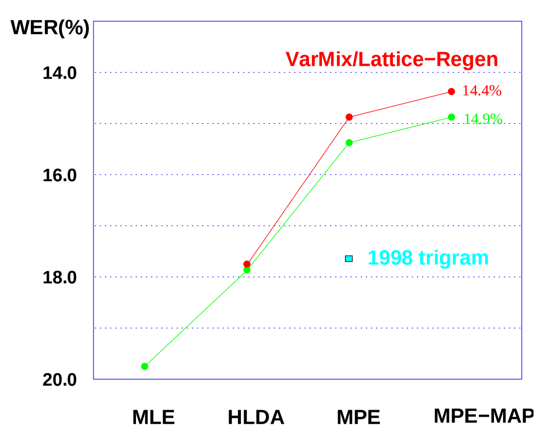
- Wordlist
  - The 59k entry wordlist was chosen from BN LM training texts according to weighted sum of frequencies to minimize the OOV rate on BNdev03
  - 0.47% OOV rate on BNdev03
- Word-based language models
  - Good-Turing discounting with the HTK HLM toolkit on sets A, B, and E
  - Modified Kneser-Ney discounting with SRI toolkit on small sets C and D
  - All models merged into a single model
  - Entropy-based pruning
  - Pruned model has 8.8M bigrams, 12.7M trigrams, and 6.6M fourgrams

Language Model (III)

- Class-based trigram
  - Trained on broadcast material (sets A, B, C, and D) with HTK HLM
  - 1,000 automatically derived classes based on word bigram statistics
- Interpolation of word-based models with class-based trigram
  - The resulting word-based model was interpolated with the class-based model with weights of (0.87:0.13)
  - The interpolation weights were computed using EM
- Perplexities on BNdev03 with word-based trigram, fourgram, and interpolated fourgram with class-based trigram are 140.9, 121.5, and 119.1 respectively.

Improved Acoustic Model + New LM: Results on BNdev03

- Marginal gain by VarMix
- VarMix/Lattice-Regeneration approach showed consistent gain over previous MPE models



%WER of various acoustic models on BNdev03 with new LM(trigram)

SAT

	MPE-MAP+HLDA	SAT	SAT-VarMix
1-best MLLR	14.1	13.4	13.4
lat-MLLR 2trans	13.8	13.5	13.4
lat-MLLR 2trans+FV	13.6	13.3	13.0

%WER of SAT models on BNdev03

Note: All the experimental results here were obtained with an preliminary version of 2003 lanuage model(fg). Since we had WB SAT model only, NB results from MPE-MAP+HLDA 1-best MLLR was used to calculate %WER

- Show specific, gender-dependent clustering for test data
- SAT training used constrained MLLR
  - one transform for silence, another for speech
  - 5 iterations of interleaved transform and MLE model update
  - 6 iterations of MPE training with fixed transform

SPron

- Single Pronunciation dictionary
- Choose one pronunciation variant based on alignment of the training data
- Same approach as in CTS
- 6919 clustered states, 16 Gaussians/state, context dependent triphone gender-dependent / bandwidth-dependent acoustic modeling
- Acoustic model was built same way as MPron (MLE→HLDA→VarMix→Lattice-Regen-MPE→Lattice-Regen-MPE-MAP)
- Final GD SPron outperforms GD MPron by 0.5% abs. on BNdev03

Conclusions

- Successfully ported many techniques from CTS to BN
- Effective discriminative GD acoustic modeling using MPE-MAP
- Improved MPE performance by Lattice-Regeneration
- SAT: successful combination with MPE on BN
- SPron outpeforms MPron

CU-HTK Fast System Description

Gunnar Evermann, Do Yeong Kim, Lan Wang, Phil Woodland
+ Rest of the HTK STT team

May 19th 2003

Cambridge University Engineering Department

Overview

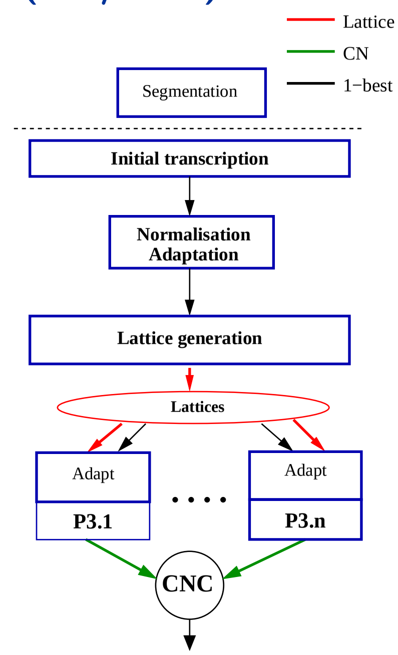
- Introduction
- System structure for 10xRT
- Review of previous 10xRT CU-HTK systems
- 10xRT system development
- 2003 system results
- Conclusions

Introduction

- Recently increased interest in making state-of-the-art eval systems fast and thus feasible for practical use
- Several sites have had systems for 10xRT BN and unlimited CTS for some time (Primary condition for RT02)
- RT04/05 will be much more difficult with limits on CTS and <5xRT BN
- CTS is harder, due to higher task & system complexity
- Prepare for future evals and concentrate on appropriate techniques
- Build and submit prototype systems (10xRT CTS in RT02 & RT03)

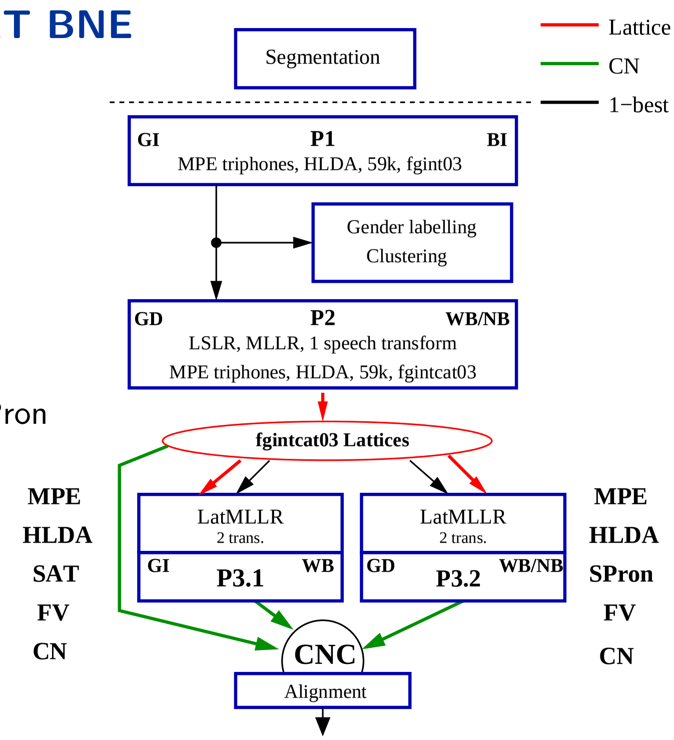
General system structure for 10xRT (BN/CTS)

- Segmentation
- Initial transcription **1xRT**
- Normalisation (re-segment, VTLN, etc.)
  Adaptation **0.5xRT**
- Lattice generation with word+class LM **4xRT**
- Lattice rescoring: for each model set: **2xRT**
  - Adaptation: MLLR (1-best + lattice), FV
  - Lattice rescoring
  - Confusion network generation
- System combination



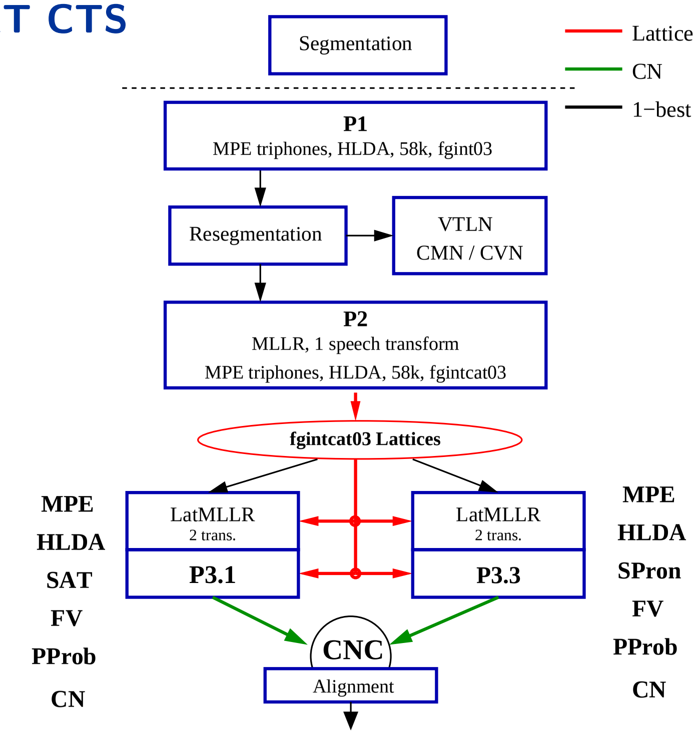
2003 System structure 10xRT BNE

- Automatic segmentation
- Speaker clustering
- All models use MPE, HLDA
- P2:gender-/bandwidth-specific MPron
- P3:
  - SAT for wideband
  - SPron for M/F and NB/WB
- 3-way system combination



2003 System structure 10xRT CTS

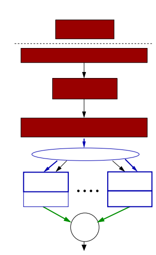
- Automatic segmentation
- Use new models from full system
- All models use MPE, HLDA
- P2: MPron models for latgen
- Use lattice MLLR and full-variance
- Selected most effective 2-way combination (SAT & SPron)



Previous work

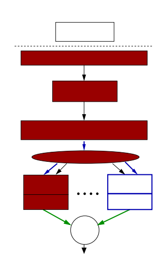
10xRT 1998 BN CUHTK-Entropic system:

- Single branch, two pass system, no lattice rescoring
- Automatic segmentation, speaker clustering
- Purpose-built acoustic models



10xRT 2002 CTS CUHTK system:

- Simple three pass system, built in a few of days based on full 320xRT system.
- Used models from full system (incl. 4 year old Pass 1 models!)
- No system combination



How to make it run fast

- All decoding parameters were carefully chosen to stay in compute budget
- Important to limit worst-case behaviour (max model beams, lattice pruning)
- Simplify adaptation, e.g. use 2 speech transforms instead of 4
- Buy many fast computers! For eval and, more importantly, experiments. CUED compute infrastructure:
  - cluster of IBM x335 dual Xeons
  - SunGrid batch queuing system (400k jobs since Nov'02)
  - for eval runs: keep all data local, use 20 fastest single CPUs (2.8GHz) turn around for 6 hour CTS set: 3 hours
- Avoid excessive overhead (e.g. reading LMs) by running on large subsets, e.g. complete BN shows or sets of several CTS sides

CTS: Development results on eval02

	Swbd1	Swbd2	Cellular	Total
P1	28.7	36.3	40.2	35.5
P2	22.4	26.8	29.8	26.6
P3.1-cn	20.4	24.0	26.1	23.7
P3.3-cn	20.4	24.3	26.6	24.0
final	19.9	23.5	25.8	23.3

%WER on eval02 (automatic segmentation) for 2003 10xRT system

- The system ran in 9.17 xRT
- The confidence scores have an NCE of 0.295

CTS: Final results on eval03

	Swbd	Fisher	Total
P1	39.0	29.7	34.5
P2	29.4	20.9	25.3
P3.1-cn	26.0	18.8	22.5
P3.3-cn	26.3	18.9	22.7
final	25.5	18.4	22.1

%WER on eval03 for 2003 10xRT system

- The system ran in 9.21 xRT
- The confidence scores have an NCE of 0.318

CTS: Progress over last year

CUED internal aims were:

- Automate running of 10xRT system
- Outperform last year's full 320xRT system in 10xRT
- Narrow gap between full and fast systems

	Swbd1	Swbd2	Cellular	Total	fast gap
320xRT 2002†	19.8	24.3	27.0	23.9	
10xRT 2002†	22.3	27.7	31.0	27.2	+14%
190xRT 2003	18.6	22.3	23.7	21.7	
10xRT 2003	19.9	23.5	25.8	23.3	+7%

%WER on eval02 for full and fast systems

†: using manual segmentation

gap on eval03 is 7%, on the progress set it is 5%.

BN: Development results on bndev03

	WER
P1	15.9
P2.fgintcat	13.1
P2.fgintcat-cn	12.8
P3.1-cn†	12.0
P3.3-cn	12.1
final	11.6

%WER on bndev03 for 2003 10xRT system
† wideband only, narrowband from P3.3

- The confidence scores have an NCE of 0.393

BN: Final results on eval03

	WER
P1	**14.6**
P2.fgintcat	11.9
P2.fgintcat-cn	11.6
P3.1-cn†	11.4
P3.3-cn	11.4
final	**10.7**

%WER on eval03 for 2003 10xRT system
† wideband only, narrowband from P3.3

- P1 ran in 0.88 xRT – submited as contrast, not an optimised 1xRT system!
- The full system ran in 9.10 xRT
- The confidence scores have an NCE of 0.412

BN: System combination

- Combination in BN system is more complicated than CTS, as we had no BN narrow-band SAT models
- Employ 3-way combination (P2, SAT, SPron) for wideband, 2-way (P2, SPron) otherwise.
- Mismatch of posterior distributions due to lattice sizes (P2 are much bigger than P3)
- Ongoing work: Investigate mapped posteriors, system weights etc.

Conclusions

- BN: rebuilt setup and constructed state-of-the-art 10xRT system
- CTS: good improvements over RT02 systems
- Narrowed gap between 100+ xRT and 10xRT considerably
- Infrastructure for quick-turnaround *system* tests (vs. single *model* experiments)

Future Work

- Optimise models (HMMs and LMs) for fast systems
- Fast versions of VTLN and MLLR
- Adaptive optimisation of decoding parameters & structure

CU-HTK RT03 Mandarin CTS System

Bin Jia, Khe Chai Sim, Mark Gales, Thomas Hain,
Andrew Liu, Phil Woodland, Kai Yu & the HTK STT Team

May 19th 2003

Cambridge University Engineering Department

Mandarin CTS 2003 System

- Acoustic and Language Model Training Data
- Mandarin Phone Sets
- Tonal Decision Tree Questions
- Vocal Tract Length Normalisation and Pitch
- Varmix and MPE training
- Results

Acoustic Training Set-Up

- Acoustic/Training Test Data:
  - training data: 34.9 hours, 379 sides, from LDC CallHome (22.4hrs) and CallFriend (12.5hrs), 451K Words (+7K English word), 628K Characters
  - development data: dev02 1.94 hours from CallFriend
- Front-end
  - Reduced bandwidth 125–3800 Hz
  - 12 PLP cepstral parameters + C0 and 1st/2nd derivatives
  - Side-based cepstral mean and variance normalisation
  - Optional vocal tract length normalisation in training and test
  - Optional pitch (and derivatives) obtained from ESPS
- Acoustic Models
  - Gender independent models
  - Decision tree state clustered, context dependent triphones
  - Approximately 3000 distinct states

Language Model

- Sources of data (using LDC character-to-word segmentor)
  - Acoustic training data (modifier Kneser-Ney)
  - News corpora: TDT[2,3,4], China Radio, People's Daily, Xinhua (Good-Turing)
- Word LMs - 11K vocabulary, 0.17% OOV on dev02

Data	Bigram	Trigram
Acoustic	206.6	190.8
Acoustic+News Corpora	199.6	179.8

Perplexity results on dev02

- Class-based LM - 75 classes trained on acoustic trnascriptions

LM	Bigram	Trigram
Class	196.1	190.1
Class+Word	188.3	172.1

Perplexity results on dev02

Mandarin Phone Sets

# Phone Set	CER (%)
59-phone	58.1
46-phone	57.0

%CER for dev02 using 12 mix comp VTLN MLE trained systems and word trigram LM

- Two phone sets considered:
  - 59-phone set, start with LDC 60 phone set, remove tone markers and
    u:e → ue
  - 46-phone set, start with 59-phone set and split long final phones, e.g.
    [aeiu]n → [aeiu] n
    [aeio]ng → [aeio] ng
    uang → ua ng
- Mapping reduced CER by 1.1% absolute
- 46 phone set was used for all further experiments

Tonal Decision Tree Questions

Tonal Questions	CER (%)
×	57.0
√	55.7

%CER for dev02 using 12 mix comp VTLN MLE trained systems and word trigram LM

- Tonal questions incorporated into decision tree process (without pitch features):
  - 3% of possible questions were tonal
  - all tonal questions used for at least one tree
  - tonal questions normally used near top of decision tree
- Yields about 1.3% absolute reduction in character error rate
- Tonal questions were used for all further experiments

VTLN/Pitch Results

VTLN	Pitch	12 Comp	+HLDA	+Pitch
×	×	57.5	56.1	—
×	√	57.0	56.2	—
√	×	55.7	53.8	—
√	√	54.6	53.4	53.0

%CER for dev02 using MLE trained systems and word-trigram LM

- HLDA used to project from static/1st/2nd/3rd derivatives to 39 dim
- Normalised pitch extracted using ESPS
  (+Pitch static/1st/2nd derivatives appended *after* HLDA)
- Results:
  - VTLN yields 1.5%-1.8% absolute reduction in CER
  - HLDA yields 0.8%-1.9% absolute reduction in CER
  - Pitch generally useful
- VTLN was used for all further experiments

Feature Vector Dimensionality

# Dim	HLDA		
	PLP	+Pitch	Pitch
39	53.8	—	53.4
42	53.7	53.0	53.2
45	53.8	53.3	53.1
48	—	53.4	53.1

%CER for dev02 using 12 mix comp MLE trained systems and word trigram LM

- Three systems examined:
  - PLP: baseline frontend with no pitch
  - +Pitch: baseline system with pitch added *after* HLDA
  - Pitch: HLDA projection from baseline frontend *and* pitch
- Small variation in performance with dimensionality
- Consistent gain ($\approx 0.5\%$) with using pitch in addition to HLDA

Additional Mixture Components/Varmix/MPE

# Comp	MLE	+Varmix	+MPE
12	53.0	52.2	49.8
16	52.3	51.7	49.9

%CER for dev02 using HLDA +Pitch trained systems and word trigram LM

- Varmix yields 0.6%-0.8% absolute reduction in error rate
- MPE yields 2.4% absolute for 12 component system
- 16 component system MLE systems better and MPE system about same
- Too many Guassians per hour for 16 comp MPE system!

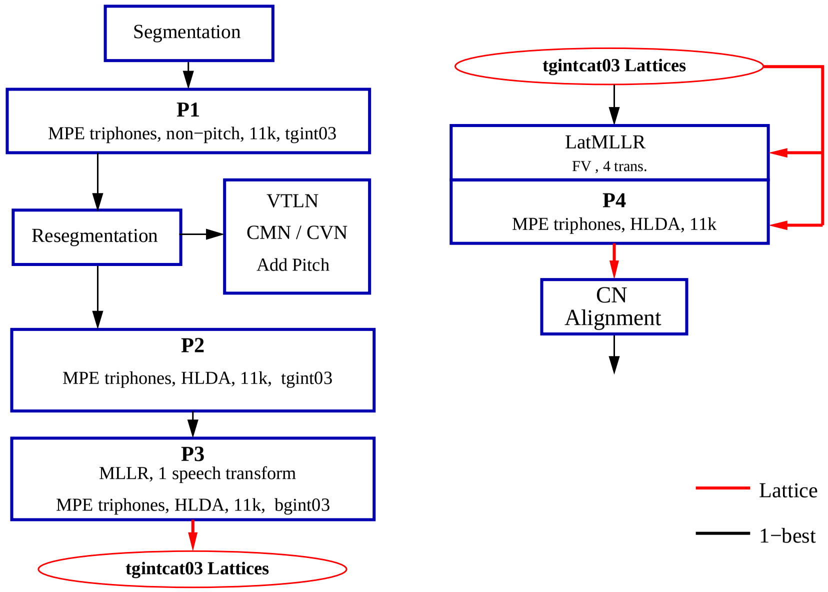
Automatic Segmentation

Segmentation	Diarisation			CER (%)
	MS	FA	Tot	
Manual	1.2	24.5	25.6	49.8
Automatic	3.7	8.3	11.9	50.8

%CER for dev02 using 12 mix comp HLDA +Pitch +Varmix MPE trained systems

- GMM classifier:
  - PLP with energy and channel energy difference plus 1st/2nd derivatives
  - 64 components for speech, 1024 components for silence.
- Diarisation score (% frame error) missed speech (MS), false alarm (FA):
  - reference derived from forced alignment of transcribed portions
  - untranscribed portions not scored (Manual MS score attribute of smoothing)
  - Manual segmentation error dominated by additional silence
- Automatic segmentation degraded CER by 1% absolute.

Mandarin RT03 System Overview



- Single system - currently no system combination.

Complete System Results

		CER (%)	
		dev02	eval03
P1	trans for VTLN	55.1	54.7
P2	trans for MLLR	50.8	51.3
P3	lat gen (bg)	49.3	50.5
	tgintcat rescore	48.9	49.8
P4	lat MLLR	48.6	49.5
CN	P4	47.9	48.6

%CER on dev02 and eval03 for all stages of 2003 system

- Final confidence scores have NCE 0.190 on eval03

Absolute Gains: dev02 vs eval03

Change to		ΔCER (%)	
		dev02	eval03
59-phone	46-phone	-1.1	-1.0
non-Tonal	Tonal	-1.3	-1.7
non-VTLN	VTLN	-1.8	-1.9
non-pitch	pitch	-1.1	-0.1
non-HLDA	HLDA	-1.9	-0.9

%CER changes on dev02 & eval03 using 12 comp MLE trained systems and word trigram LM

- dev02 numbers use manual segmentation, eval03 uses automatic segmentation
- Comparison of dev02 and eval03 gains:
  - all design choices gave improvements on both test sets
  - absolute gains differ (particularly pitch and HLDA), decisions affected by train/test speaker overlap?

Conclusions

- Current system:
  - 46 phone set, with tonal decision tree questions
  - 3 emitting states per phone model
  - VTLN, pitch, MPE and linear adaptation
  - standard techniques yield gains (but consistently less than expected)
- Future work:
  - investigate limited gains from standard schemes
  - additional systems, SAT etc, and system combination
  - alternative phone sets
  - modify HMM topology
  - add degree of voicing to frontend

Overall Conclusions

- Progress for UL CTS: same structure as 2002 + automatic segmentation and improved models

- Progress for BN: ported and verified working many techniques to BN with good results

- Fast Systems built for BN and CTS
  - Very similar architecture for BN and CTS 10xRT operation
  - CTS 10xRT system about 1% poorer than full system

- Initial Mandarin CTS system built: reasonable performance but still some way to go ...